

Docket No.: 31869-202101
(PATENT)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Tetsuji Nakagawa

Application No.: Not Yet Assigned

Confirmation No.:

Filed: Concurrently Herewith

Art Unit: N/A

For: MORPHOLOGICAL ANALYZER,
MORPHOLOGICAL ANALYSIS METHOD,
AND MORPHOLOGICAL ANALYSIS
PROGRAM

Examiner: Not Yet Assigned

CLAIM FOR PRIORITY AND SUBMISSION OF DOCUMENTS

MS Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

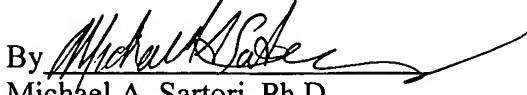
Applicant hereby claims priority under 35 U.S.C. 119 based on the following prior foreign application filed in the following foreign country on the date indicated:

Country	Application No.	Date
Japan	2003-154625	May 30, 2003

In support of this claim, a certified copy of the said original foreign application is filed herewith.

Dated: March 30, 2004

Respectfully submitted,

By 
Michael A. Sartori, Ph.D.

Registration No.: 41,289
VENABLE LLP
P.O. Box 34385
Washington, DC 20043-9998
(202) 344-4000
(202) 344-8300 (Fax)
Attorney/Agent For Applicant

DC2-534294

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2003年 5月30日
Date of Application:

出願番号 特願2003-154625
Application Number:

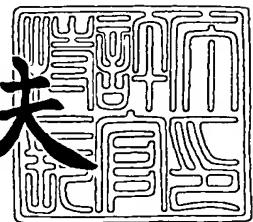
[ST. 10/C] : [JP 2003-154625]

出願人 沖電気工業株式会社
Applicant(s):

2003年11月28日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫



【書類名】 特許願
【整理番号】 KN002592
【提出日】 平成15年 5月30日
【あて先】 特許庁長官 太田 信一郎 殿
【国際特許分類】 G06F 17/27
【発明者】
【住所又は居所】 東京都港区虎ノ門1丁目7番12号 沖電気工業株式会社内
【氏名】 中川 哲治
【特許出願人】
【識別番号】 000000295
【氏名又は名称】 沖電気工業株式会社
【代表者】 篠塚 勝正
【代理人】
【識別番号】 100090620
【弁理士】
【氏名又は名称】 工藤 宣幸
【その他】 国等の委託研究の成果に係る特許出願（平成14年度通信・放送機構「多言語標準文書処理システムの研究開発」委託研究、産業活力再生特別措置法第30条の適用を受けるもの）
【手数料の表示】
【予納台帳番号】 013664
【納付金額】 21,000円
【提出物件の目録】
【物件名】 明細書 1
【物件名】 図面 1
【物件名】 要約書 1
【包括委任状番号】 9006358

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 形態素解析装置、形態素解析方法及び形態素解析プログラム

【特許請求の範囲】

【請求項 1】 形態素解析対象文に対して所定の形態素解析方法を適用し、活用形がある品詞についてはその活用形の情報を含む品詞タグが付与された単語列でなる、形態素解析結果の候補である仮説を 1 又は複数生成する仮説生成手段と、

品詞に関する複数種類の n - g r a m 確率モデルの情報を格納しているモデル格納手段と、

上記各仮説に対し、大量の文中でその仮説が出現するであろう生成確率を、上記モデル格納手段に格納されている複数種類の n - g r a m 確率モデルの情報を重み付けて結合して求める生成確率計算手段と、

上記各仮説の生成確率に基づき、解となる仮説を探索する解探索手段とを備え、

上記モデル格納手段は、少なくとも、品詞及び品詞の活用形を反映させた種類の n - g r a m 確率モデルの情報を格納している

ことを特徴とする形態素解析装置。

【請求項 2】 品詞及び品詞の活用形を反映させた種類の n - g r a m 確率モデルの情報を、上記仮説の i 番目の単語 ω_i の品詞を t_i 、品詞の活用形を t_i form としたとき、その品詞 t_i をとる単語の中で単語 ω_i が出現する条件付き確率 $P(\omega_i | t_i)$ と、単語 ω_i に係る品詞 t_i POS がその活用形 t_i form で出現する条件付き確率 $P(t_i$ form $| t_i$ POS) と、直前 $N-1$ 個の単語に係る品詞タグ列 $t_{i-N+1} \cdots t_{i-1}$ の並びに続いて単語 ω_i に係る品詞 t_i POS が出現する条件付き確率 $P(t_i$ POS $| t_{i-N+1} \cdots t_{i-1}$) との積になっていることを特徴とする請求項 1 に記載の形態素解析装置。

【請求項 3】 上記モデル格納手段は、複数種類の 1 種類として、クラス n - g r a m 確率モデルの情報をも格納していることを特徴とする請求項 1 又は 2 に

記載の形態素解析装置。

【請求項4】 クラスn-gram確率モデルの情報におけるクラスの種別は、品詞タグ付きコーパスと品詞タグ無しコーパスとから学習したものであることを特徴とする請求項3に記載の形態素解析装置。

【請求項5】 形態素解析対象文に対して所定の形態素解析方法を適用し、活用形がある品詞についてはその活用形の情報を含む品詞タグが付与された単語列でなる、形態素解析結果の候補である仮説を1又は複数生成する仮説生成工程と、

上記各仮説に対し、大量の文中でその仮説が出現するであろう生成確率を、予め用意されている、品詞及び品詞の活用形を反映させた種類のn-gram確率モデルの情報を含む、品詞に関する複数種類のn-gram確率モデルの情報を重み付けて結合して求める生成確率計算工程と、

上記各仮説の生成確率に基づき、解となる仮説を探索する解探索工程とを含むことを特徴とする形態素解析方法。

【請求項6】 請求項5の形態素解析方法を、コンピュータが実行可能なコードで記述していることを特徴とする形態素解析プログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は形態素解析装置、形態素解析方法及び形態素解析プログラムに関し、特に、複数の正解候補の中から最適な解を高い精度で選択し得るようにしたものである。

【0002】

【従来の技術】

形態素解析装置は、入力された文に対してその文を構成する各形態素を同定して区切り、品詞を付与するものである。しかしながら、形態素に分割する際及び品詞を付与する際に、複数の正解候補が存在し曖昧性が発生するため、正解候補の中から正しいものを選択する必要がある。

【0003】

このような目的のために、以下のような品詞n-gramモデルに基づく方法がいくつか提案されている。

【0004】

【特許文献1】 特開平7-271792号公報

【0005】

【非特許文献1】 浅原、松本著、形態素解析のための拡張統計モデル」、情処論Vol. 43, No. 3, pp. 685-695, 2002

特許文献1は、日本語形態素解析において、統計的手法によりこの曖昧性を解決する方法について述べている。直前の2つの品詞が与えられたときの3つ目の品詞が出現する確率である品詞三つ組確率と、品詞が与えられたときの単語の出現確率である品詞別単語出力確率から、文を構成する単語列と各単語に付与された品詞列の同時確率を最大にするような候補を選ぶことにより、曖昧性の解消を行っている。

【0006】

非特許文献1では、特徴的な性質を持つ形態素の品詞を語彙化し、似た性質を持つ品詞をグループ化するという拡張を行うことで、より精度の高い形態素解析を実現している。

【0007】

【発明が解決しようとする課題】

しかしながら、特許文献1の記載方法は、過去の品詞系列のみから次に来る品詞を予測し、さらに品詞が与えられた場合の条件のみから単語を予測しているため、高い精度で形態素解析を行うのは困難である。つまり、助詞等の機能語はしばしば他の形態素と異なる特徴的な性質をもつが、このような語に関しては品詞だけではなく語彙自体の情報も考慮する必要がある。また、品詞体系によっては数百を越える数の品詞を扱わなければならないこともあるが、そのような場合は品詞の組合せの数が膨大になるため、特許文献1の記載方法を直接適用して形態素解析を行うことは困難である。

【0008】

非特許文献1の記載方法では、品詞の語彙化により特徴的な性質を持つ形態素

に対処している。また、品詞のグループ化を行うことにより品詞の数が多い場合にも対処している。しかしながら、語彙化やグループ化は誤り駆動に基づく方法を用いて一部の形態素や品詞に関してのみ行われるため、形態素に関する十分な情報を利用できているわけではなく、また、訓練データを効果的に利用できないという課題がある。

【0009】

そのため、複数の正解候補の中から最適な解を高い精度で選択し得る形態素解析装置、形態素解析方法及び形態素解析プログラムが望まれている。

【0010】

【課題を解決するための手段】

かかる課題を解決するため、第1の本発明の形態素解析装置は、（1）形態素解析対象文に対して所定の形態素解析方法を適用し、活用形がある品詞についてはその活用形の情報を含む品詞タグが付与された単語列でなる、形態素解析結果の候補である仮説を1又は複数生成する仮説生成手段と、（2）品詞に関する複数種類のn-gram確率モデルの情報を格納しているモデル格納手段と、（3）上記各仮説に対し、大量の文中でその仮説が出現するであろう生成確率を、上記モデル格納手段に格納されている複数種類のn-gram確率モデルの情報を重み付けて結合して求める生成確率計算手段と、（4）上記各仮説の生成確率に基づき、解となる仮説を探索する解探索手段とを備え、（2-1）上記モデル格納手段が、少なくとも、品詞及び品詞の活用形を反映させた種類のn-gram確率モデルの情報を格納していることを特徴とする。

【0011】

第2の本発明の形態素解析方法は、（1）形態素解析対象文に対して所定の形態素解析方法を適用し、活用形がある品詞についてはその活用形の情報を含む品詞タグが付与された単語列でなる、形態素解析結果の候補である仮説を1又は複数生成する仮説生成工程と、（2）上記各仮説に対し、大量の文中でその仮説が出現するであろう生成確率を、予め用意されている、品詞及び品詞の活用形を反映させた種類のn-gram確率モデルの情報を含む、品詞に関する複数種類のn-gram確率モデルの情報を重み付けて結合して求める生成確率計算工程と

、（3）上記各仮説の生成確率に基づき、解となる仮説を探索する解探索工程とを含むことを特徴とする。

【0012】

第2の本発明の形態素解析プログラムは、第2の本発明の形態素解析方法を、コンピュータが実行可能なコードで記述していることを特徴とする。

【0013】

【発明の実施の形態】

（A）第1の実施形態

以下、本発明による形態素解析装置、形態素解析方法及び形態素解析プログラムの第1の実施形態を図面を参照しながら説明する。

【0014】

（A-1）第1の実施形態の構成

図1は、第1の実施形態の形態素解析装置の機能的構成を示すブロック図である。第1の実施形態の形態素解析装置は、例えば、入出力装置や補助記憶装置などを備えるパソコン等の情報処理装置上に、形態素解析プログラム（図2～図4参照）をインストールすることによって実現されるが、機能的には、図1で表すことができる。

【0015】

第1の実施形態の形態素解析装置100は、大きくは、確率モデルを使用して形態素解析を行う解析部110、確率モデル等を格納するモデル格納部120、及び、パラメータ学習用のコーパスから確率的モデルの学習を行うためのモデル学習部130から構成されている。

【0016】

解析部110は、形態素解析を行う文を入力するための入力部111、入力された文に対して、形態素辞書格納部121に格納されている形態素辞書を用いて可能な解（形態素解析結果）の候補（仮説）を生成する仮説生成部112、生成された各仮説に対して、確率モデル格納部122に格納された品詞n-g r a mモデル、語彙化品詞n-g r a mモデル（当該モデルの定義については後述する）及び階層化品詞n-g r a mモデル（当該モデルの定義については後述する）

を、重み格納部123に格納された重み付けにより結合して生成確率を計算する生成確率計算部113、生成確率の付与された仮説の中から最も尤度の高い解を選ぶ解探索部114、及び、解探索部114により得られた解を出力する出力部115より構成される。

【0017】

なお、入力部111は、例えば、キーボード等の一般的な入力部だけでなく、記録媒体のアクセス装置等のファイル読込装置や、文書をイメージデータとして読み込んでそれをテキストデータに置き換える文字認識装置等も該当する。また、出力部115は、例えば、ディスプレイやプリンタ等の一般的な出力部だけでなく、記録媒体へ格納する記録媒体アクセス装置等も該当する。

【0018】

モデル格納部120は、確率推定部132で計算され、生成確率計算部113及び重み計算部133で使用される確率モデルを格納した確率モデル格納部122、重み計算部133で計算され、生成確率計算部113で使用される重みを格納する重み格納部123、及び、仮説生成部112で解候補（仮説）を生成するために使用される形態素辞書を格納する形態素辞書格納部121から構成されている。

【0019】

モデル学習部130は、確率推定部132及び重み計算部133でモデルの学習を行うために使用される品詞タグ付きコーパス格納部131、品詞タグ付きコーパス格納部131に格納された品詞タグ付きコーパスを用いて確率モデルの推定を行い、その結果を確率モデル格納部122へ格納する確率推定部132、及び、確率モデル格納部122に格納された確率モデルと品詞タグ付きコーパス格納部131に格納された品詞タグ付きコーパスを用いて確率モデルの重みを計算し、その結果を重み格納部123へ格納する重み計算部133から構成されている。

【0020】

（A-2）第1の実施形態の動作

次に、第1の実施形態の形態素解析装置100の動作（第1の実施形態の形態

素解析方法) を、図2のフローチャートを参照しながら説明する。図2は、入力された文を形態素解析装置100が形態素解析して出力するまでの処理の流れを示すフローチャートである。

【0021】

まず、使用者が入力した形態素解析をしたい文を入力部111によって取り込む(201)。入力された文に対して、仮説生成部112は、形態素辞書格納部121に格納された形態素辞書を用いて、可能な解の候補である仮説を生成する(202)。この仮説生成部112による処理は、例えば、一般的な形態素解析方法を適用する。生成確率計算部113は、確率モデル格納部122及び重み格納部123に格納された情報を用いて、仮説生成部112で生成された各仮説に対しその生成確率を計算する(203)。生成確率計算部113は、各仮説に対する生成確率として、品詞n-gram、語彙化品詞n-gram及び階層化品詞n-gramを確率的に重み付けたものを計算する。

【0022】

ここで、入力された文の先頭から($i + 1$)番目の単語及びその品詞タグをそれぞれ ω_i 及び t_i とし、文中の単語(形態素)の数をnとする。また、品詞タグ t は、品詞 t POSと活用形 t formからなっているとする。なお、活用形がない品詞の場合には、品詞と品詞タグとは同一のものである。仮説、つまり正解候補の単語・品詞タグ列は、

$$\omega_0 t_0 \dots \omega_{n-1} t_{n-1}$$

と表現され、このような仮説の中から最も生成確率の高いものを解として選べばよいため、(1)式を満足する最適な単語・品詞タグ列を求める事になる。

【0023】

例えば、「私は見た。」という文章は、「私(名詞；より細かく分類した代名詞を適用しても良い)／は(助詞；より細かく分類した副助詞を適用しても良い)／見(動詞-連用形)／た(助動詞)／。(句点)」という単語・品詞タグ列と、「私(名詞)／は(助詞)／見(動詞-終止形)／た(助動詞)／。(句点)」という単語・品詞タグ列との2つの仮説が生じ、いずれが最適であるかが(1)式によって求められる。なお、この例の場合、「見」に関してのみ、「動詞

」という品詞と「連用形」又は「終止形」という活用形で品詞タグが構成され、他の単語（句点も1個の単語として取扱う）については品詞のみで品詞タグが構成されている。

【0024】

【数1】

$$\begin{aligned}
 & \hat{w}_0 \hat{t}_0 \cdots \hat{w}_{n-1} \hat{t}_{n-1} \\
 = & \underset{w_0 t_0 \cdots w_{n-1} t_{n-1}}{\operatorname{argmax}} P(w_0 t_0 \cdots w_{n-1} t_{n-1}) \\
 = & \underset{w_0 t_0 \cdots w_{n-1} t_{n-1}}{\operatorname{argmax}} \prod_{i=0}^{n-1} P(w_i t_i | w_0 t_0 \cdots w_{i-1} t_{i-1}) \\
 = & \underset{w_0 t_0 \cdots w_{n-1} t_{n-1}}{\operatorname{argmax}} \prod_{i=0}^{n-1} \sum_{M \in \mathbf{M}} P(M | w_0 t_0 \cdots w_{i-1} t_{i-1}) P(w_i t_i | w_0 t_0 \cdots w_{i-1} t_{i-1} M) \quad (1)
 \end{aligned}$$

$$\begin{aligned}
 \mathbf{M} = & \{M_{\text{POS}}^1, \dots, M_{\text{POS}}^{N_{\text{POS}}}, \\
 & M_{\text{lex1}}^1, \dots, M_{\text{lex1}}^{N_{\text{lex1}}}, M_{\text{lex2}}^1, \dots, M_{\text{lex2}}^{N_{\text{lex2}}}, M_{\text{lex3}}^1, \dots, M_{\text{lex3}}^{N_{\text{lex3}}}, \\
 & M_{\text{hier}}^1, \dots, M_{\text{hier}}^{N_{\text{hier}}}\} \quad (2)
 \end{aligned}$$

$$\sum_{M \in \mathbf{M}} P(M) = 1 \quad (2.5)$$

(1) 式において、第1行の「 $\hat{w}_0 \hat{t}_0 \cdots \hat{w}_{n-1} \hat{t}_{n-1}$ 」は最適な単語・品詞タグ列を意味しており、 argmax は、複数の単語・品詞タグ列（仮説）の中から生成確率 $P(w_0 t_0 \cdots w_{n-1} t_{n-1})$ が最も高い単語・品詞タグ列を選択することを表している。

【0025】

ある単語・品詞タグ列の生成確率 $P(w_0 t_0 \cdots w_{n-1} t_{n-1})$ は、その単語・品詞タグ列においてその $(i+1)$ 番目（ i は $0 \sim (n-1)$ ）の単語・品詞タグが生じる条件付き確率 $P(w_i t_i | w_0 t_0 \cdots w_{n-1} t_{n-1})$ の積で表される。条件付き確率 $P(w_i t_i | w_0 t_0 \cdots w_{n-1} t_{n-1})$ は、ある n -gramモデル M で計算される単語についての出力確率 $P(w_i t_i | w_0 t_0 \cdots w_{n-1} t_{n-1} M)$ と、その n -gramモデル M に対する重み $P(M | w_0 t_0 \cdots w_{n-1} t_{n-1})$ との積を、全てのモデルについて求めた積和で表される。

【0026】

ここで、出力確率 $P(\omega_i t_i | \omega_0 t_0 \dots \omega_{n-1} t_{n-1} M)$ の情報が、確率モデル格納部122に格納されており、 $n\text{-gram}$ モデルMに対する重み $P(M | \omega_0 t_0 \dots \omega_{n-1} t_{n-1})$ の情報が、重み格納部123に格納されていいる。

【0027】

(2) 式は、生成確率 $P(\omega_0 t_0 \dots \omega_{n-1} t_{n-1})$ の計算に適用される全てのモデルMを集合Mとして記載したものである。但し、集合Mは、(2.5)式に示すように、その要素である各モデルM毎の確率 $P(M)$ が1になるようなモデルの集合である。

【0028】

モデルMについての下付パラメータはモデルの種類を表しており、「POS」は品詞 $n\text{-gram}$ モデルを表しており、「lex1」は第1の語彙化品詞 $n\text{-gram}$ モデルを表しており、「lex2」は第2の語彙化品詞 $n\text{-gram}$ モデルを表しており、「lex3」は第3の語彙化品詞 $n\text{-gram}$ モデルを表しており、「hier」は階層化品詞 $n\text{-gram}$ モデルを表している。モデルMについての上付パラメータは、そのモデルにおける記憶長の長さ $N-1$ 、言い換えると、 $n\text{-gram}$ での単語数（品詞タグ数も同数）を表している。

【0029】

【数2】

 $\mathcal{M}_{\text{POS}}^N$: 品詞 N -gram モデル

$$P(w_i t_i | w_0 t_0 \cdots w_{i-1} t_{i-1} \mathcal{M}_{\text{POS}}^N) \equiv P(w_i | t_i) P(t_i | t_{i-N+1} \cdots t_{i-1}) \quad (3)$$

 $\mathcal{M}_{\text{lex1}}^N, \mathcal{M}_{\text{lex2}}^N, \mathcal{M}_{\text{lex3}}^N$: 語彙化品詞 N -gram モデル

$$P(w_i t_i | w_0 t_0 \cdots w_{i-1} t_{i-1} \mathcal{M}_{\text{lex1}}^N) \equiv P(w_i | t_i) P(t_i | w_{i-N+1} t_{i-N+1} \cdots w_{i-1} t_{i-1}) \quad (4)$$

$$P(w_i t_i | w_0 t_0 \cdots w_{i-1} t_{i-1} \mathcal{M}_{\text{lex2}}^N) \equiv P(w_i | t_i | t_{i-N+1} \cdots t_{i-1}) \quad (5)$$

$$P(w_i t_i | w_0 t_0 \cdots w_{i-1} t_{i-1} \mathcal{M}_{\text{lex3}}^N) \equiv P(w_i | t_i | w_{i-N+1} t_{i-N+1} \cdots w_{i-1} t_{i-1}) \quad (6)$$

 $\mathcal{M}_{\text{hier}}^N$: 階層化品詞 N -gram モデル

$$P(w_i t_i | w_0 t_0 \cdots w_{i-1} t_{i-1} \mathcal{M}_{\text{hier}}^N) \equiv P(w_i | t_i) P(t_i^{\text{form}} | t_i^{\text{POS}}) P(t_i^{\text{POS}} | t_{i-N+1} \cdots t_{i-1}) \quad (7)$$

記憶長の長さ $N - 1$ の品詞 $n - g r a m$ モデルは、(3) 式で定義される。記憶長の長さ $N - 1$ の品詞 $n - g r a m$ モデルは、品詞タグ t_i をとる中でその単語 w_i が出現する条件付き確率 $P(w_i | t_i)$ と、直前 $N - 1$ 個の単語に係る品詞タグ列 $t_{i-N+1} \cdots t_{i-1}$ の並びに続いてその単語 w_i の品詞タグ t_i が出現する条件付き確率 $P(t_i | t_{i-N+1} \cdots t_{i-1})$ との積で定義される。

【0030】

記憶長の長さ $N - 1$ の第 1 の語彙化品詞 $n - g r a m$ モデルは、(4) 式で定義される。記憶長の長さ $N - 1$ の第 1 の語彙化品詞 $n - g r a m$ モデルは、品詞タグ t_i をとる中でその単語 w_i が出現する条件付き確率 $P(w_i | t_i)$ と、直前 $N - 1$ 個の単語・品詞タグ列 $w_{i-N+1} t_{i-N+1} \cdots w_{i-1} t_{i-1}$ の並びに続いて、その単語 w_i の品詞タグ t_i が出現する条件付き確率 $P(t_i | w_{i-N+1} t_{i-N+1} \cdots w_{i-1} t_{i-1})$ との積で定義される。

【0031】

記憶長の長さ $N - 1$ の第 2 の語彙化品詞 $n - g r a m$ モデルは、(5) 式で定義される。記憶長の長さ $N - 1$ の第 2 の語彙化品詞 $n - g r a m$ モデルは、直前 $N - 1$ 個の単語に係る品詞タグ列 $t_{i-N+1} \cdots t_{i-1}$ の並びに続いて、単語 w_i とその品詞タグ t_i との組み合わせ $w_i t_i$ が出現する条件付き確率 $P(w_i | t_i | t_{i-N+1} \cdots t_{i-1})$ との積で定義される。

$\omega_i t_i | \omega_{i-1} t_{i-1} \cdots \omega_1 t_1$ で定義される。

【0032】

記憶長の長さ $N-1$ の第3の語彙化品詞 $n-gram$ モデルは、(6) 式で定義される。記憶長の長さ $N-1$ の第3の語彙化品詞 $n-gram$ モデルは、直前 $N-1$ 個の単語・品詞タグ列 $\omega_{i-N+1} t_{i-N+1} \cdots \omega_{i-1} t_{i-1}$ の並びに続いて、単語 ω_i とその品詞タグ t_i との組み合わせ $\omega_i t_i$ が出現する条件付き確率 $P(\omega_i t_i | \omega_{i-1} t_{i-1} \cdots \omega_1 t_1)$ で定義される。

【0033】

記憶長の長さ $N-1$ の階層化品詞 $n-gram$ モデルは、(7) 式で定義される。記憶長の長さ $N-1$ の階層化品詞 $n-gram$ モデルは、その品詞 t_i をとる単語の中で候補単語 ω_i が出現する条件付き確率 $P(\omega_i | t_i)$ と、単語 ω_i に係る品詞 t_i POS がその活用形 $t_i form$ で出現する条件付き確率 $P(t_i form | t_i POS)$ と、直前 $N-1$ 個の単語に係る品詞タグ列 $t_{i-N+1} t_{i-N} \cdots t_{i-1}$ の並びに続いて単語 ω_i に係る品詞 t_i POS が出現する条件付き確率 $P(t_i POS | t_{i-N+1} t_{i-N} \cdots t_{i-1})$ の積で定義される。なお、単語 ω_i に係る品詞 t_i POS がその活用形 $t_i form$ で出現する条件付き確率 $P(t_i form | t_i POS)$ は、活用形が存在しない品詞については常に「1」として取扱う。

【0034】

生成確率計算部 113 によって、各仮説に対する生成確率 $P(\omega_0 t_0 \cdots \omega_{n-1} t_{n-1})$ が計算されると、解探索部 114 は、(1) 式に示すように、その中で最も生成確率が高い解を選択する(図2の204)。

【0035】

上述したように、生成確率計算部 113 による、各仮説に対する生成確率 $P(\omega_0 t_0 \cdots \omega_{n-1} t_{n-1})$ の計算を行った後に、解探索部 114 による最も生成確率が高い解(最適解)の探索を行っても良いが、例えば、ビタビ(Viterbi)アルゴリズムを適用して、生成確率計算部 113 による処理と、解探索部 114 による処理とを融合して行うようにしても良い。すなわち、入力

された文の先頭から $(i + 1)$ 番目までの単語・品詞タグ列を規定するパラメータ i を徐々に大きくしながら行う、ビタビアルゴリズムによる最適な単語・品詞タグ列の探索によって、生成確率計算部 113 による処理と、解探索部 114 による処理とを融合して行って、最適解を探索する。

【0036】

上述した (1) 式を満足する最適解の単語・品詞タグ列が求まると、出力部 115 によって、求まった最適解 (形態素解析結果) をユーザへ出力する (205)。

【0037】

次に、モデル学習部 130 の動作、すなわち、生成確率計算部 113 において使用する確率モデル及び確率モデルの重みを、予め用意された品詞タグ付きコーパスから計算して求める動作を、図 3 を参照しながら説明する。

【0038】

まず、確率推定部 132 により、以下に示す確率モデルのパラメータを学習する (301)。

【0039】

ここで、単語列、品詞列、品詞タグ列、及び又は、単語・品詞タグ列などの系列を X とし、その系列 X が品詞タグ付きコーパス格納部 131 に格納されたコーパス中に出現した回数を $f(X)$ で表すと、各確率モデルに対するパラメータは、以下のように表される。

【0040】

【数3】

 $\mathcal{M}_{\text{POS}}^N$: 品詞 N -gram モデル

$$P(w_i|t_i) = \frac{f(t_i w_i)}{f(t_i)} \quad (8)$$

$$P(t_i|t_{i-N+1} \dots t_{i-1}) = \frac{f(t_{i-N+1} \dots t_{i-1} t_i)}{f(t_{i-N+1} \dots t_{i-1})} \quad (9)$$

 $\mathcal{M}_{\text{lex1}}^N, \mathcal{M}_{\text{lex2}}^N, \mathcal{M}_{\text{lex3}}^N$: 語彙化品詞 N -gram モデル

$$P(w_i|t_i) = \frac{f(t_i w_i)}{f(t_i)} \quad (10)$$

$$P(t_i|w_{i-N+1} t_{i-N+1} \dots w_{i-1} t_{i-1}) = \frac{f(w_{i-N+1} t_{i-N+1} \dots w_{i-1} t_{i-1} t_i)}{f(w_{i-N+1} t_{i-N+1} \dots w_{i-1} t_{i-1})} \quad (11)$$

$$P(w_i t_i|t_{i-N+1} \dots t_{i-1}) = \frac{f(t_{i-N+1} \dots t_{i-1} w_i t_i)}{f(t_{i-N+1} \dots t_{i-1})} \quad (12)$$

$$P(w_i t_i|w_{i-N+1} t_{i-N+1} \dots w_{i-1} t_{i-1}) = \frac{f(w_{i-N+1} t_{i-N+1} \dots w_{i-1} t_{i-1} w_i t_i)}{f(w_{i-N+1} t_{i-N+1} \dots w_{i-1} t_{i-1})} \quad (13)$$

 $\mathcal{M}_{\text{hier}}^N$: 階層化品詞 N -gram モデル

$$P(w_i|t_i) = \frac{f(t_i^{\text{POS}} w_i)}{f(t_i^{\text{POS}})} \quad (14)$$

$$P(t_i^{\text{form}}|t_i^{\text{POS}}) = \frac{f(t_i^{\text{POS}} t_i^{\text{form}})}{f(t_i^{\text{POS}})} \quad (15)$$

$$P(t_i^{\text{POS}}|t_{i-N+1} \dots t_{i-1}) = \frac{f(t_{i-N+1} \dots t_{i-1} t_i^{\text{POS}})}{f(t_{i-N+1} \dots t_{i-1})} \quad (16)$$

記憶長の長さ $N - 1$ の品詞 $n - g r a m$ モデルは、上述したように、(3) 式で表されるので、(3) 式の右辺の各要素 $P(w_i | t_i)$ 及び $P(t_i | t_{i-N+1} \dots t_{i-1})$ を、(8) 式及び(9) 式に従ってパラメータとして得る。

【0041】

また、記憶長の長さ $N - 1$ の第 1 ~ 第 3 の語彙化品詞 $n - g r a m$ モデルは、上述したように、(4) 式 ~ (6) 式で表されるので、(4) 式 ~ (6) 式の右辺の各要素 $P(w_i | t_i)$ 、 $P(t_i | w_{i-N+1} t_{i-N+1} \dots w_{i-1} t_{i-1})$ 、 $P(w_i t_i | t_{i-N+1} \dots t_{i-1})$ 及び $P(w_i t_i | w_{i-N+1} t_{i-N+1} \dots w_{i-1} t_{i-1})$ を、(10) 式 ~ (13) 式に従ってパラメータとして得る。

【0042】

さらに、記憶長の長さ $N - 1$ の階層化品詞 $n - g r a m$ モデルは、上述したように、(7) 式で表されるので、(7) 式の右辺の各要素 $P(w_i | t_i)$ 、 $P(t_i form | t_i POS)$ 及び $P(t_i POS | t_i - N + 1 \cdots t_i - 1)$ を、(14) 式～(16) 式に従ってパラメータとして得る。

【0043】

いずれのパラメータも、コーパス中に、該当する単語列、品詞列、品詞タグ列などが出現在した回数を数え上げ、その出現回数、及び又は、各式の分子となる出現回数を分母となる出現回数で除算した値を確率モデル格納部 122 へ格納する。

【0044】

図 5～図 7 は、確率モデル格納部 122 に格納された一部の確率モデルのパラメータを示す図面である。

【0045】

次に、品詞タグ付きコーパス格納部 131 に格納されている品詞タグ付きコーパスと確率モデル格納部 122 に格納された確率モデルを用いて、重み計算部 133 により、各確率モデルに対する重みの計算を行い、その結果を重み格納部 123 へ格納する(302；図 4 参照)。

【0046】

ここで、重みの計算については、(17) 式に示すように、単語・品詞タグ列に依存しない近似を行うこととする。そして、leave-one-out 法に基づいて、図 4 に示す手順で計算を行う。

【0047】

【数4】

$$P(\mathcal{M} | w_0 t_0 \cdots w_{i-1} t_{i-1}) \simeq P(\mathcal{M}) \quad (17)$$

まずははじめに、各モデル M に対する重みパラメータ $\lambda(M)$ を全て 0 にする初期化を行う(401)。次に、品詞タグ付きコーパス格納部 131 に格納されている品詞タグ付きコーパスから、単語と品詞タグの対を 1 つ取り出して $w_0 t_0$

とし、その i 個前にある単語と品詞をそれぞれ $\omega - 1$ $t - 1$ とする (402)。次に、各確率モデル M に対して確率 $P'(\omega_0 t_0 | \omega - N + 1 t - N + 1 \cdots \omega - 1 t - 1 M)$ を計算する (403)。

【0048】

ここで、確率 $P'(X | Y) = P'(\omega_0 t_0 | \omega - N + 1 t - N + 1 \cdots \omega - 1 t - 1 M)$ は、現在考慮している事象を数え上げの対象から除いて求めた確率値で、(18) 式のようにコーパス中に出現した事象の数を用いて計算する。

【0049】

【数5】

$$P'(X|Y) = \begin{cases} 0 & (f(Y) - 1 = 0) \\ \frac{f(XY) - 1}{f(Y) - 1} & \text{otherwise.} \end{cases} \quad (18)$$

以上のようにして各モデルに対し計算した確率値の中で、最も高い値を返したモデルを M' とすると、このモデルに対する重みパラメータ $\lambda(M')$ を 1 だけ増やす (404)。ステップ 402 ~ 404 でなる処理を、品詞タグ付きコーパス中の全ての単語と品詞タグとの対について繰り返し (405)、全ての単語と品詞タグとの対に対する処理が終了すると、各確率モデル M に対して、(19) 式に示す正規化した重み $P(M)$ を求める (406)。

【0050】

【数6】

$$P(M) = \frac{\lambda(M)}{\sum_N \lambda(N)} \quad (19)$$

なお、上記では、簡単のために、(17) 式のように重みの計算に近似を用いたが、かわりに品詞 $n - g r a m$ 、語彙化 $n - g r a m$ 及び階層化品詞 $n - g r a m$ 等の結合を用いて、(1) 式と同様に重みを計算することもできる。

【0051】

(A-3) 第1の実施形態の効果

上記第1の実施形態によれば、形態素辞書を利用して得た複数の形態素解析結

果（仮説）から最尤のものを決定する際に、品詞の情報に加え、品詞を語彙化した情報、及び、品詞の階層を考慮した情報を使用してその仮説の生成確率を計算して最尤なものを決定するようにしたので、品詞の情報のみを使用して生成確率を計算して最尤な仮説を決定する方法に比べ、より頑健で高精度な解析を行うことができ、曖昧性を解消できる。

【0052】

(B) 第2の実施形態

次に、本発明による形態素解析装置、形態素解析方法及び形態素解析プログラムの第1の実施形態を図面を参照しながら説明する。

【0053】

(B-1) 第2の実施形態の構成

図8は、第2の実施形態の形態素解析装置の機能的構成を示すブロック図である。第2の実施形態の形態素解析装置も、例えば、入出力装置や補助記憶装置などを備えるパソコン等の情報処理装置上に、形態素解析プログラム（図9～図11参照）をインストールすることによって実現されるが、機能的には、図8で表すことができる。

【0054】

第2の実施形態の形態素解析装置500は、大きく見た場合には、第1の実施形態の構成にクラスタリング部540が加わったものであり、また、モデル学習部530においても、第1の実施形態の構成に、品詞タグ無しコーパス格納部534及び品詞タグ・クラス付きコーパス格納部535が加わったものである。

【0055】

クラスタリング部540は、クラス学習部541、クラスタリングパラメータ格納部542及びクラス付与部543を有する。

【0056】

クラス学習部541は、品詞タグ付きコーパス格納部531中に格納されている品詞タグ付きコーパス及び品詞タグ無しコーパス格納部534に格納されている品詞タグ無しコーパスを用いてクラスの学習を行い、学習の結果得られたクラスタリング用のパラメータをクラスタリングパラメータ格納部542へ格納する

ものである。

【0057】

クラス付与部543は、クラスタリングパラメータ格納部542に格納されているクラスタリング用のパラメータを用いて、品詞タグ付きコーパス格納部531中の品詞タグ付きコーパスを入力し、これにクラスを付与したものを品詞タグ・クラス付きコーパス格納部535へ格納し、また、仮説生成部512で得られた仮説を入力し、これにクラスを付与したものを生成確率計算部513へ出力するものである。

【0058】

品詞タグ・クラス付きコーパス格納部535に格納された品詞タグ・クラス付きコーパスは、確率推定部532及び重み計算部533が利用する。

【0059】

(B-2) 第2の実施形態の動作

次に、第2の実施形態の形態素解析装置500の動作（第2の実施形態の形態素解析方法）を、図9のフローチャートを参照しながら説明する。図9は、入力された文を形態素解析装置500が形態素解析して出力するまでの処理の流れを示すフローチャートである。

【0060】

第2の実施形態の形態素解析装置500は、第1の実施形態と比べて、確率値の計算にクラス情報を用いる点だけが異なるため、以下では、第1の実施形態と異なる点についてのみ説明する。

【0061】

文の入力（601）、仮説の生成（602）が行われた後、生成された仮説をクラス付与部543へ入力してクラスの付与を行い、そのクラスが付与された仮説が生成確率計算部513に与えられる（603）。クラスの付与の方法については後述する。

【0062】

次に、クラスが付与された各仮説に対して、生成確率計算部513で生成確率の計算を行う（604）。但し、各仮説に対する生成確率は、品詞n-g r a m

、語彙化品詞 n-g r a m、階層化品詞 n-g r a m 及びクラス品詞 n-g r a m を確率的に重み付けたものを用いる。計算方法は、上述した (1) 式で表されるが、モデルの集合 M として、(2) 式に代え、次の (20) 式に示すものが適用される。但し、集合 M は、(20.5) 式に示すように、その要素である各モデル M 每の確率 $P(M)$ が 1 になるようなモデルの集合である。

【0063】

【数7】

$$\begin{aligned}
 M = & \{ M_{\text{POS}}^1, \dots, M_{\text{POS}}^{N_{\text{POS}}}, \\
 & M_{\text{lex1}}^1, \dots, M_{\text{lex1}}^{N_{\text{lex1}}}, M_{\text{lex2}}^1, \dots, M_{\text{lex2}}^{N_{\text{lex2}}}, M_{\text{lex3}}^1, \dots, M_{\text{lex3}}^{N_{\text{lex3}}}, \\
 & M_{\text{hier}}^1, \dots, M_{\text{hier}}^{N_{\text{hier}}}, \\
 & M_{\text{class1}}^1, \dots, M_{\text{class1}}^{N_{\text{class1}}}, M_{\text{class2}}^1, \dots, M_{\text{class2}}^{N_{\text{class2}}}\} \tag{20}
 \end{aligned}$$

$$\sum_{M \in M} P(M) = 1 \tag{20.5}$$

(2) 式及び (20) 式の比較から明らかのように、第 2 の実施形態においては、第 1 及び第 2 のクラス品詞 n-g r a m モデルも適用されている。

【0064】

(20) 式において、下付パラメータが「class1」のものが第 1 のクラス品詞 n-g r a m モデルを表しており、下付パラメータが「class2」のものが第 2 のクラス品詞 n-g r a m モデルを表している。

【0065】

【数8】

$M_{\text{class1}}^N, M_{\text{class2}}^N$: クラス品詞 N -gram モデル

$$P(w_i t_i | w_0 t_0 \dots w_{i-1} t_{i-1} M_{\text{class1}}^N) \equiv P(w_i | t_i) P(t_i | c_{i-N+1} t_{i-N+1} \dots c_{i-1} t_{i-1}) \tag{21}$$

$$P(w_i t_i | w_0 t_0 \dots w_{i-1} t_{i-1} M_{\text{class2}}^N) \equiv P(w_i | t_i) P(t_i | c_{i-N+1} t_{i-N+1} \dots c_{i-1} t_{i-1}) \tag{22}$$

記憶長の長さ $N-1$ の第 1 のクラス品詞 n-g r a m モデルは、(21) 式で定義され、記憶長の長さ $N-1$ の第 2 のクラス品詞 n-g r a m モデルは、(22) 式で定義される。

【0066】

記憶長の長さ $N - 1$ の第 1 のクラス品詞 $n - g r a m$ モデルは、品詞タグ t_i をとる中でその単語 ω_i が出現する条件付き確率 $P(\omega_i | t_i)$ と、直前 $N - 1$ 個の単語に係るクラス・品詞タグ列 $c_i - N + 1 t_i - N + 1 \cdots c_i - 1 t_i - 1$ の並びに続いてその単語 ω_i の品詞タグ t_i が出現する条件付き確率 $P(t_i | c_i - N + 1 t_i - N + 1 \cdots c_i - 1 t_i - 1)$ の積で定義される。

【0067】

記憶長の長さ $N - 1$ の第 2 のクラス品詞 $n - g r a m$ モデルは、直前 $N - 1$ 個のクラス・品詞タグ列 $c_i - N + 1 t_i - N + 1 \cdots c_i - 1 t_i - 1$ の並びに続いて、単語 ω_i とその品詞タグ t_i との組み合わせ $\omega_i t_i$ が出現する条件付き確率 $P(\omega_i t_i | \omega_i - N + 1 t_i - N + 1 \cdots \omega_i - 1 t_i - 1)$ で定義される。

【0068】

このようなクラスを利用して単語の出現確率を予測することにより、品詞や語彙化した品詞とは異なる情報も用いて、仮説の生成確率を計算することが可能となっている。また、クラスを用いた形態素解析方法は既に知られているが、当該形態素解析装置 500 は、上述のように、クラス品詞 $n - g r a m$ 以外の確率モデルと確率的に重み付けをして結合して用いるため、クラスを用いたことによる精度の低下等の副作用が起りにくい。

【0069】

以上のように、確率モデルにより、各仮説に対する生成確率の計算を行った後、最適解の探索を行い（605）、結果を出力する（606）。

【0070】

図 10 は、上述の生成確率計算部 513 において使用する確率モデル及び確率モデルの重みを、あらかじめ用意された品詞タグ付きコーパス及び品詞タグ無しコーパスを用いて求める処理を示すフローチャートである。

【0071】

まず、クラス学習部 541 により、品詞タグ付きコーパス格納部 531 に格納されている品詞タグ付きコーパス及び品詞タグ無しコーパス格納部 534 に格納

されている品詞タグ無しコーパスを用いて、クラスタリングのためのパラメータを学習し、クラスタリングパラメータ格納部542へ格納する（701）。

【0072】

但し、ここでのクラスタリングは、コーパス中の単語情報のみを用いて、その単語にクラスを与えるものとする。そのため、クラスタリングのパラメータの学習には、作成するのが困難な品詞タグ付きコーパスだけでなく容易に入手可能な品詞タグ無しコーパスを用いることができる。このようなクラスタリングを行う方法の一つとして、隠れマルコフモデルを用いることができ、この場合、B a u m-W e l c h アルゴリズムによりパラメータの学習を行うことができる。隠れマルコフモデルの学習及びクラスの付与については、例えば、『L. R a b i n e r, B-H. Ju a n g著、古井監訳、「音声認識の基礎（下）」、1995年』等に詳しく紹介されている。

【0073】

次に、クラスタリングパラメータ格納部542中のクラスタリング用パラメータを用いて、クラス付与部543は、品詞タグ付きコーパス格納部531に格納された品詞タグ付きコーパスを入力し、各単語のクラスタリングを行い、クラスを付与し、そのクラスの付与された品詞タグ付きコーパスを品詞タグ・クラス付きコーパス格納部535へ格納する（702）。次に、確率推定部532により、確率モデルのパラメータを学習する（703）。

【0074】

ここで、クラス品詞n-g r a mモデル以外の各確率モデルに対するパラメータは、第1の実施形態の場合と同様に学習する。単語列、品詞タグ列、クラス・品詞タグ列などの系列をXとし、その系列Xが品詞タグ・クラス付きコーパス格納部535に格納されたコーパス中に出現した回数をf(X)で表すと、クラス品詞n-g r a mモデルに対するパラメータは、（23）式～（25）式のように表される。

【0075】

【数9】

$M_{\text{class1}}^N, M_{\text{class2}}^N$: クラス品詞 N -gram モデル

$$P(w_i | t_i) = \frac{f(t_i w_i)}{f(t_i)} \quad (23)$$

$$P(t_i | c_{i-N+1} t_{i-N+1} \cdots c_{i-1} t_{i-1}) = \frac{f(c_{i-N+1} t_{i-N+1} \cdots c_{i-1} t_{i-1} t_i)}{f(c_{i-N+1} t_{i-N+1} \cdots c_{i-1} t_{i-1})} \quad (24)$$

$$P(w_i t_i | c_{i-N+1} t_{i-N+1} \cdots c_{i-1} t_{i-1}) = \frac{f(c_{i-N+1} t_{i-N+1} \cdots c_{i-1} t_{i-1} w_i t_i)}{f(c_{i-N+1} t_{i-N+1} \cdots c_{i-1} t_{i-1})} \quad (25)$$

記憶長の長さ $N - 1$ の第1及び第2のクラス品詞 $n - g r a m$ モデルは、上述したように、(21)及び(22)式で表されるので、(21)式及び(22)式の右辺の各要素 $P(w_i | t_i)$ 、 $P(t_i | c_{i-N+1} t_{i-N+1} \cdots c_{i-1} t_{i-1})$ 及び $P(w_i t_i | c_{i-N+1} t_{i-N+1} \cdots c_{i-1} t_{i-1})$ を、(23)式～(25)式に従ってパラメータとして得る。

【0076】

各確率モデルでのパラメータを確率モデル格納部522へ格納した後には、重み計算部533において重みの計算を行い、その結果を重み格納部523へ格納する(704)。

【0077】

重みの計算については、図11のフローチャートに示す手順で行う。第2の実施形態の重みの計算も、品詞タグ付きコーパス格納部131に格納されている品詞タグ付きコーパスの代わりに品詞タグ・クラス付きコーパス格納部535に格納されている品詞タグ・クラス付きコーパスを利用する点、品詞 $n - g r a m$ 、語彙化品詞 $n - g r a m$ 及び階層化品詞 $n - g r a m$ に加えて、クラス品詞 $n - g r a m$ を確率モデルとして用いる点を除けば、第1の実施形態の重み計算の処理(図4参照)と同様であるので、その処理の詳細説明は省略する。

【0078】

(B-3) 第2の実施形態の効果

上記第2の実施形態によれば、形態素辞書を利用して得た複数の形態素解析結果(仮説)から最尤のものを決定する際に、クラスタリングにより付与したクラス情報をも用いるようにしたので、品詞よりは細かく、語彙化した品詞よりは抽

象化された情報を利用でき、より頑健で高精度な解析を行うことができる。また、品詞タグ無しデータを利用してクラスタリングの精度を高めているので、形態素解析結果の精度も高まっている。

【0079】

(C) 他の実施形態

上記第1の実施形態では、仮説の生成確率を、品詞 $n - g r a m$ 確率モデル、語彙化品詞 $n - g r a m$ 確率モデル及び階層化品詞 $n - g r a m$ 確率モデルを利用して求めるものを示し、第2の実施形態では、仮説の生成確率を、品詞 $n - g r a m$ 確率モデル、語彙化品詞 $n - g r a m$ 確率モデル、階層化品詞 $n - g r a m$ 確率モデル及びクラス品詞 $n - g r a m$ 確率モデルを利用して求めるものを示したが、本発明は、適用する複数種類の確率モデルの中に階層化品詞 $n - g r a m$ 確率モデルが含まれていれば、複数種類の確率モデルの組み合わせは、上記実施形態のものに限定されない。

【0080】

また、仮説生成部112、512による仮説（形態素解析結果候補）の生成方法は、形態素辞書を利用した一般的な形態素解析方法に限定されず、文字に関する $n - g r a m$ を利用した形態素解析方法など、他の形態素解析方法を利用するようにしても良い。

【0081】

さらに、上記各実施形態では、最尤の仮説である形態素解析結果を出力するものを示したが、得られた形態素解析結果を、機械翻訳部などの自然言語処理部に直ちに与えるようにしても良い。

【0082】

さらにまた、上記各実施形態では、モデル学習部やクラスタリング部を備えるものを示したが、モデル学習部やクラスタリング部を備えないで、解析部とモデル格納部とで形態素解析装置を構成するようにしても良い。この場合、モデル格納部への情報は、予めモデル学習部やクラスタリング部で形成されたものである。また、第2の実施形態でクラスタリング部などを省略した場合には、モデル格納部にクラス付与機能を持たせることを要する。

【0083】

また、各種の処理に供するコーパスは、通信処理により、ネットワークなどから取り込むようなものであっても良い。

【0084】

本発明が適用可能な言語は、上記実施形態のような日本語には限定されないことは勿論である。

【0085】**【発明の効果】**

以上のように、本発明によれば、複数の正解候補の中から最適な解を高い精度で選択し得る形態素解析装置、形態素解析方法及び形態素解析プログラムを提供できる。

【図面の簡単な説明】**【図1】**

第1の実施形態の形態素解析装置の機能的構成を示すブロック図である。

【図2】

第1の実施形態の形態素解析装置の解析時動作を示すフローチャートである。

【図3】

第1の実施形態の形態素解析装置のモデル学習動作を示すフローチャートである。

【図4】

図3の重みの計算処理の詳細を示すフローチャートである。

【図5】

第1の実施形態のモデルパラメータの例を示す説明図（その1）である。

【図6】

第1の実施形態のモデルパラメータの例を示す説明図（その2）である。

【図7】

第1の実施形態のモデルパラメータの例を示す説明図（その3）である。

【図8】

第2の実施形態の形態素解析装置の機能的構成を示すブロック図である。

【図9】

第2の実施形態の形態素解析装置の解析時動作を示すフローチャートである。

【図10】

第2の実施形態の形態素解析装置のモデル学習動作を示すフローチャートである。

【図11】

図10の重みの計算処理の詳細を示すフローチャートである。

【符号の説明】

100、500…形態素解析装置、

110、510…解析部、

112、512…仮説生成部、113、513…生成確率計算部、

114、514…解探索部、

120、520…モデル格納部、

121、521…形態素辞書格納部、122、522…確率モデル格納部、

123、523…重み格納部、

130、530…モデル学習部、

131、531…品詞タグ付きコーパス格納部、

132、532…確率推定部、133、533…重み計算部、

534…品詞タグ無しコーパス格納部、

535…品詞タグ・クラス付きコーパス格納部、

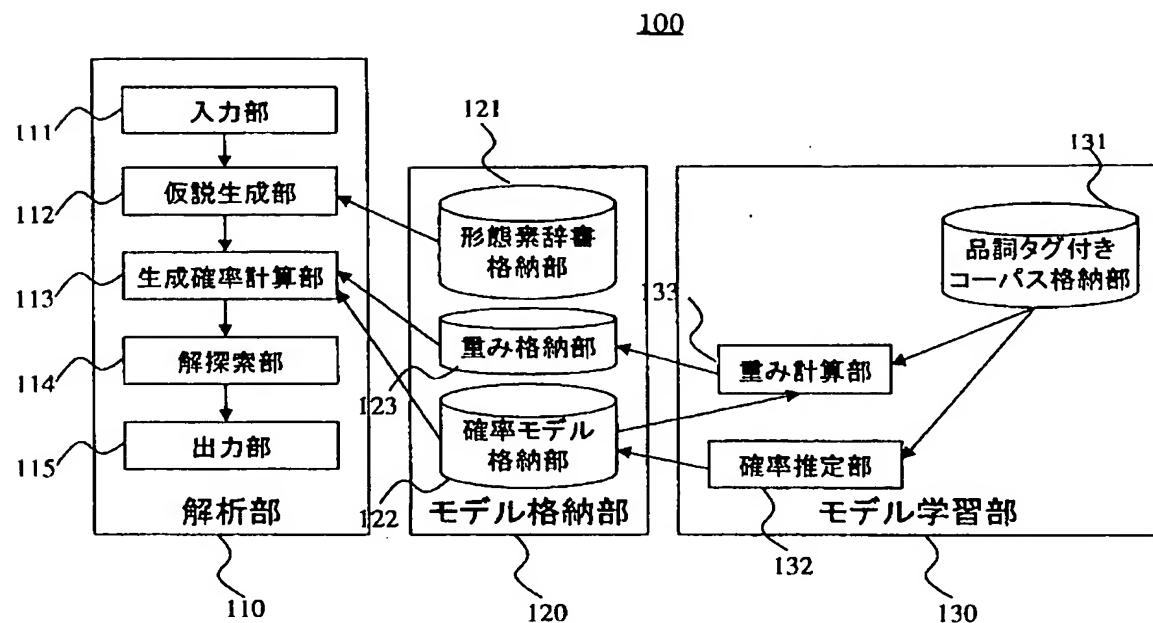
540…クラスタリング部、

541…クラス学習部、542…クラスタリングパラメータ格納部、

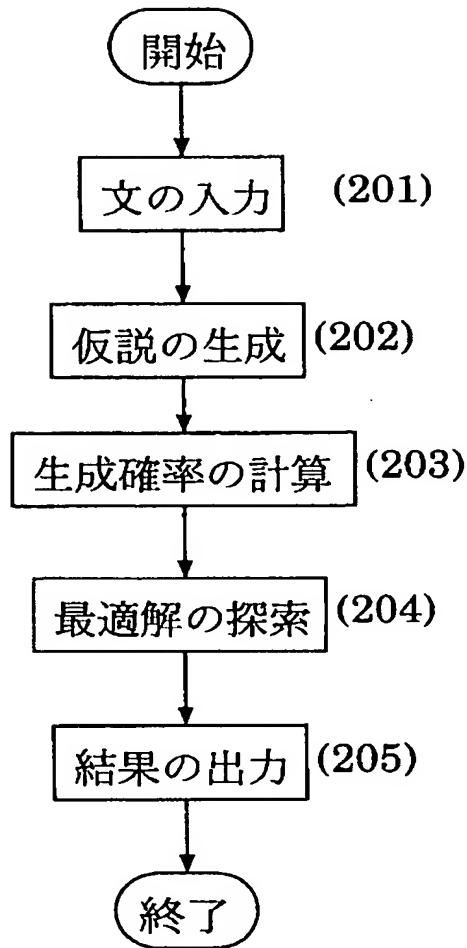
543…クラス付与部。

【書類名】 図面

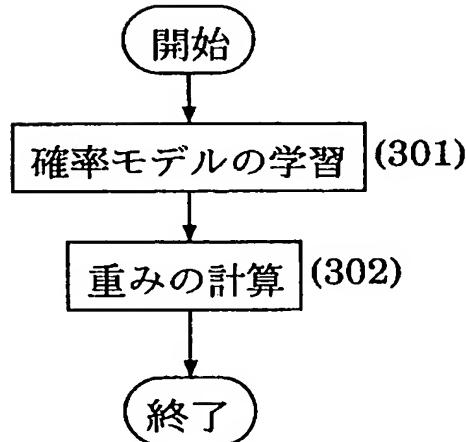
【図 1】



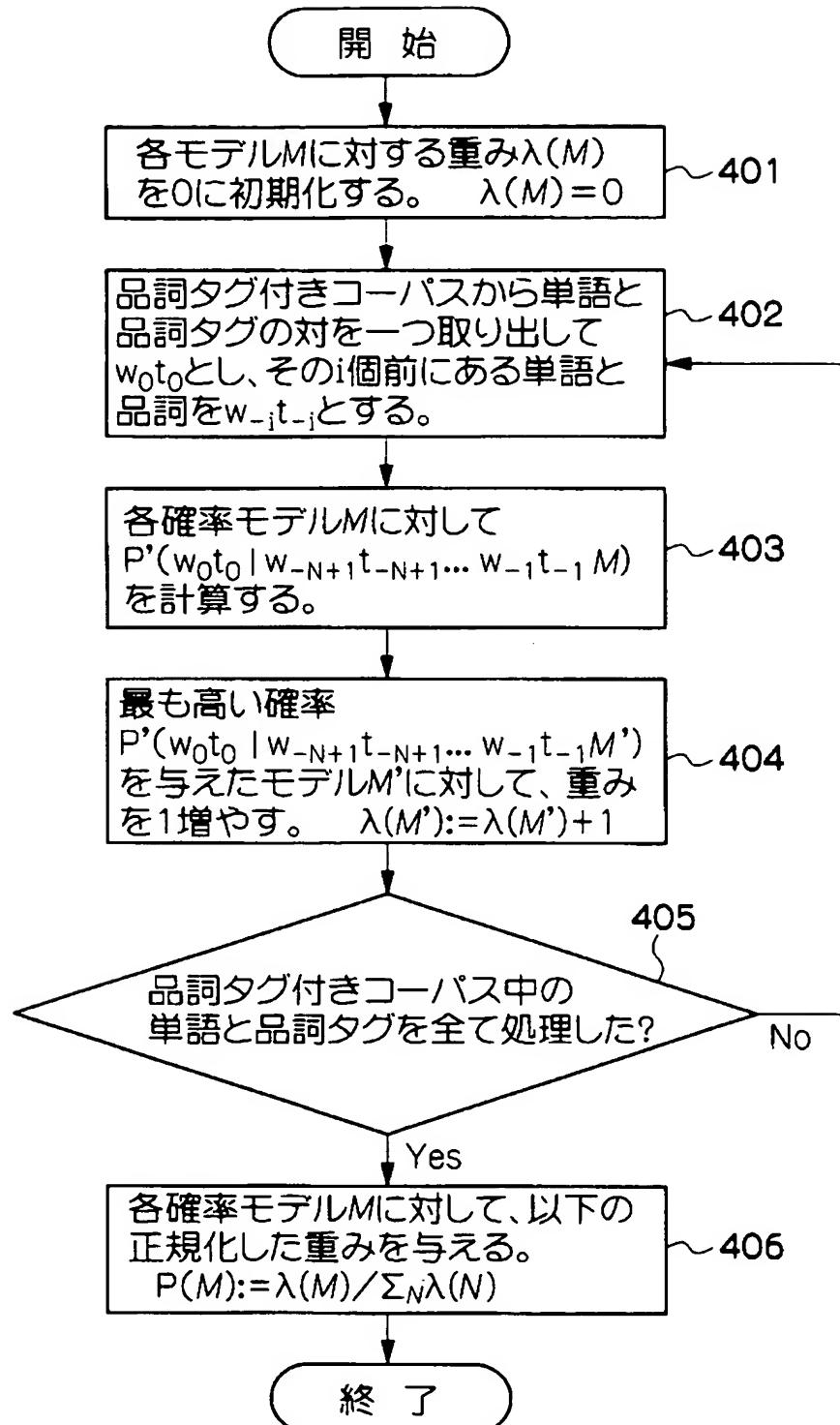
【図2】



【図3】



【図4】



【図 5】

P(政権 | 名詞-普通名詞)=0.004343
P(裁判官 | 名詞-普通名詞)=0.000328
P(中国語 | 名詞-普通名詞)=0.000036
P(コート | 名詞-普通名詞)=0.000018
P(倫理 | 名詞-普通名詞)=0.000182
P(騒然と | 形容詞:基本連用形)=0.001216
P(悠長な | 形容詞:ダ列基本連体形)=0.000515
P(手渡さ | 動詞:未然形)=0.000319
P(損なわ | 動詞:未然形)=0.000559
P(結集し | 動詞:基本連用形)=0.000178
•
•
•

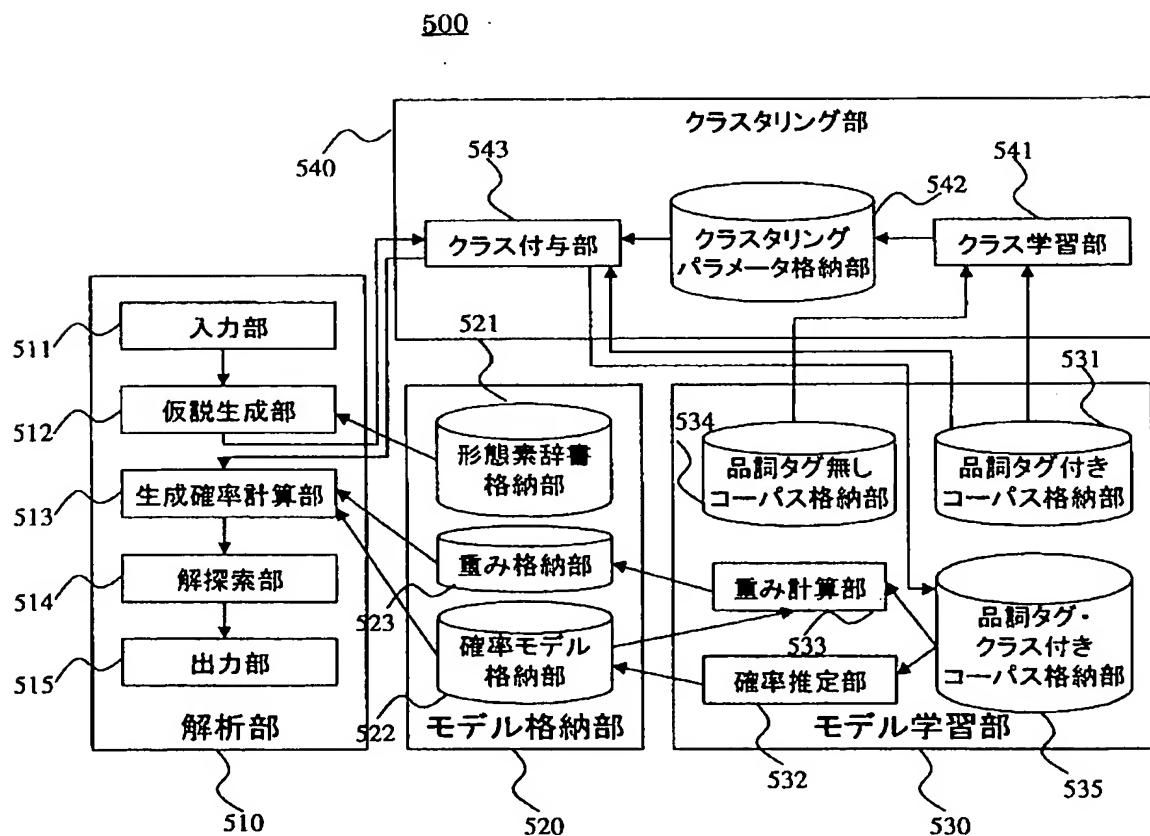
【図 6】

P(名詞-普通名詞 | 名詞-数詞)=0.075674
P(副詞 | 副詞)=0.012787
P(特殊-句点 | 助詞-終助詞)=0.734557
P(名詞-固有名詞 | 副詞)=0.000314
P(動詞:未然形 | 副詞)=0.025575
P(形容詞:基本形 | 副詞)=0.033498
P(名詞-サ変名詞 | 副詞)=0.129285
P(動詞:基本形 | 副詞)=0.034832
P(特殊-句点 | 形容詞:基本推量形)=0.857143
P(特殊-読点 | 動詞:基本連用形)=0.543240
•
•
•

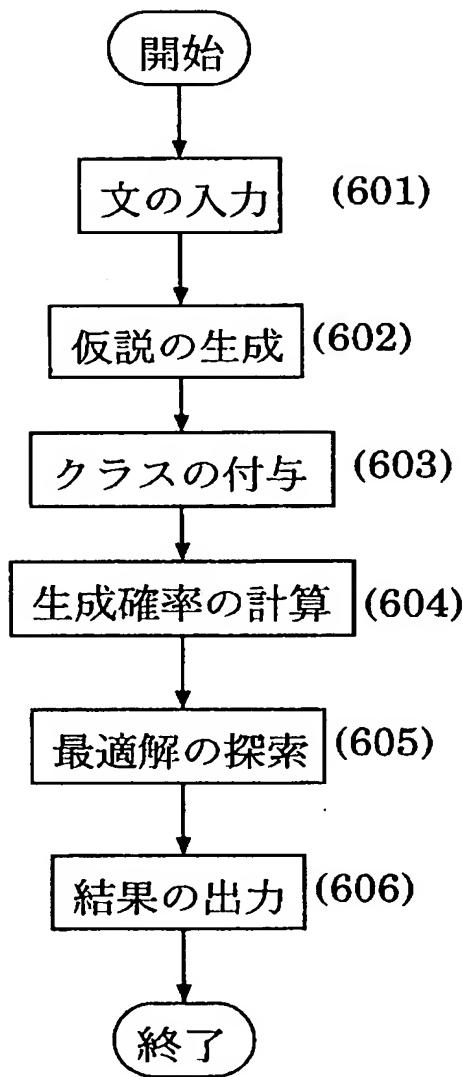
【図 7】

$P(\text{。}_\text{特殊-句点} | \text{つかんだ}_\text{動詞:夕形}) = 0.562500$
 $P(\text{に}_\text{助詞-格助詞} | \text{代わり}_\text{名詞-普通名詞}) = 0.782609$
 $P(\text{多方}_\text{名詞-人名} | \text{は}_\text{助詞-副助詞}) = 0.000032$
 $P(\text{した}_\text{動詞:夕形} | \text{録音}_\text{名詞-サ変名詞}) = 0.285714$
 $P(\text{が}_\text{助詞-格助詞} | \text{漁船}_\text{名詞-普通名詞}) = 0.166667$
 $P(\text{不必要だ}_\text{形容詞:基本形} | \text{全く}_\text{副詞}) = 0.010638$
 $P(\text{市場}_\text{名詞-普通名詞} | \text{海外}_\text{名詞-普通名詞}) = 0.021858$
 $P(\text{教室}_\text{名詞-普通名詞} | \text{で}_\text{助詞-格助詞}) = 0.000154$
 $P(\text{は}_\text{助詞-副助詞} | \text{まで}_\text{助詞-接続助詞}) = 0.060000$
 $P(\text{を}_\text{助詞-格助詞} | \text{業務}_\text{名詞-普通名詞}) = 0.461538$
 .
 .
 .

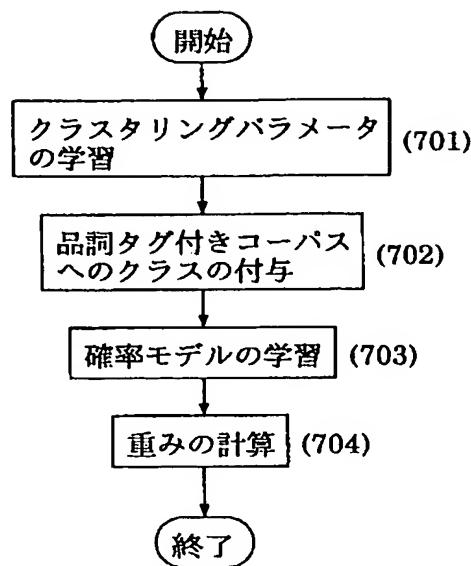
【図 8】



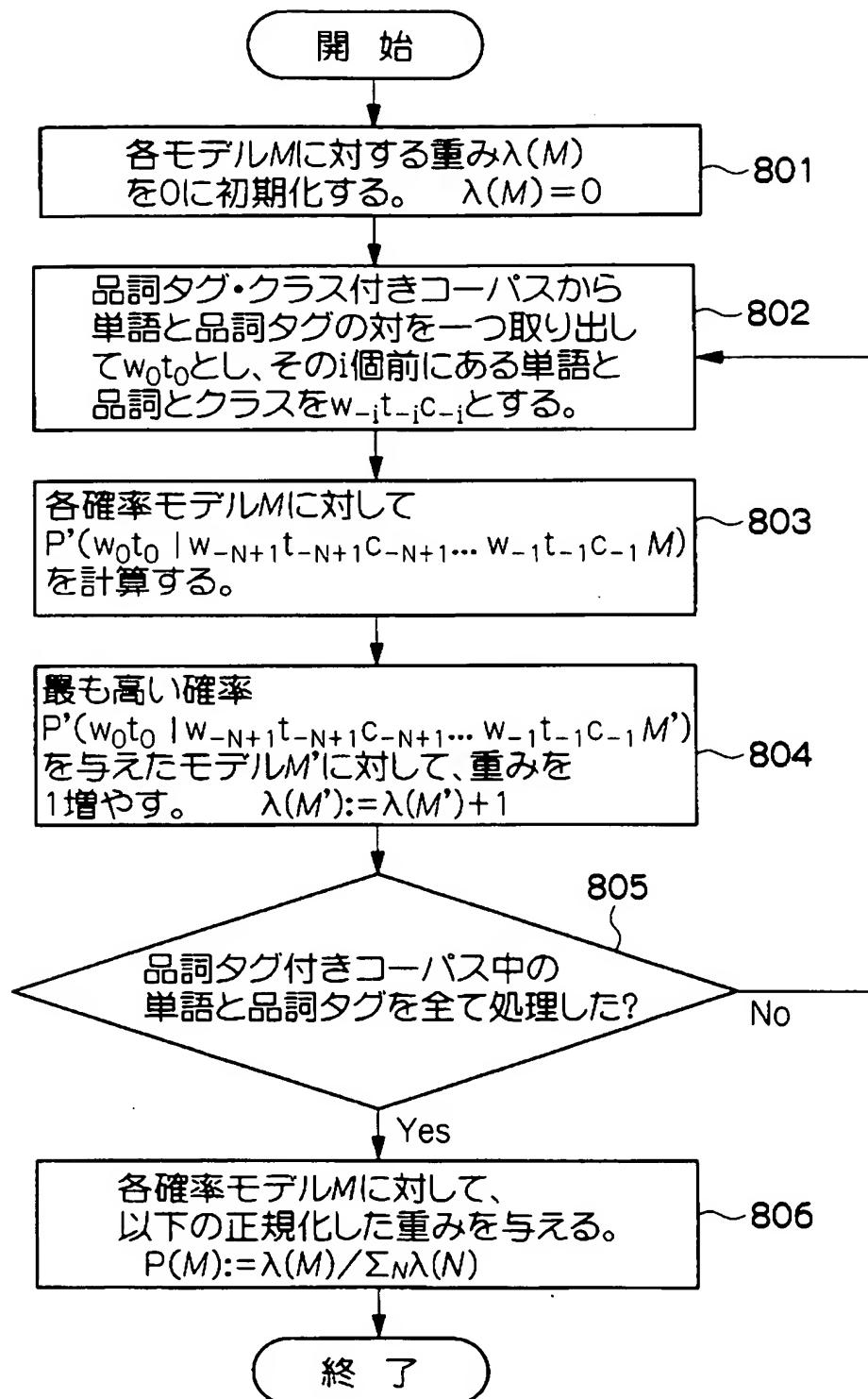
【図9】



【図10】



【図11】



【書類名】 要約書

【要約】

【課題】 複数の正解候補の中から最適な解を高い精度で選択し得る形態素解析装置、形態素解析方法及び形態素解析プログラムを提供する。

【解決手段】 本発明では、文を形態素解析して得た形態素解析結果の候補である各仮説に対し、品詞 $n - g r a m$ 確率モデル、語彙化した品詞に係る語彙化品詞 $n - g r a m$ 確率モデル、及び、品詞を品詞本体と品詞の活用形とに階層化して求められている階層化品詞 $n - g r a m$ 確率モデルを重み付け結合して、生成確率を求め、各仮説の生成確率に基づいて、解（最終的な形態素解析結果）を探索する。

【選択図】 図 1

特願2003-154625

出願人履歴情報

識別番号 [000000295]

1. 変更年月日 1990年 8月22日

[変更理由] 新規登録

住 所 東京都港区虎ノ門1丁目7番12号

氏 名 沖電気工業株式会社